一个面向原始数据搜寻的快速射电暴数据集

徐志骏[1,2*]，安涛[1,2]，郭绍光[1,2]，劳保强[1,2]，吕唯佳[1,2]，伍筱聪[1,2]

1. 中国科学院上海天文台 SKA 区域中心联合实验室, 上海 200030

2. 鹏城实验室 SKA 区域中心联合实验室，深圳 518066

* 联系人, E-mail: xuthus@shao.ac.cn

收稿日期: ； 接受日期: ；

国家自然科学基金资助项目（批准号：12041301）、国家重点研发计划大科学装置前沿研究专项 (2018YFA0404603)、国家自然科学基金资助项目（批准号：11873079）和中国科学院青年创新促进会项目 (编号:2021258) 资助项目。

摘要　快速射电暴是目前国际天文学新兴前沿热点，随着海量观测数据带来的处理和分析的挑战，亟需开展快速射电暴信号智能搜寻和甄别的研究。为了加速快速射电暴搜寻研究，我们开发了一套基于机器学习的快速射电暴数据集，它可以训练机器学习算法以搜寻原始数据中的快速射电暴。目前数据集有 8020 个快速射电暴仿真图像、4010 个非快速射电暴和 4010 个射频干扰仿真图像，这些图像是根据开放的快速射电暴观测结果构建的，并可根据需要扩展数量。本研究旨在为最先进的人工智能算法提供开源数据集，以测试和比较快速射电暴识别算法。该数据集为卷积神经网络和经典机器学习算法提供图像和 numpy 格式的文件。数据集可以实现快速射电暴和非快速射电暴分类，或快速射电暴、射频干扰和背景噪声分类。在本例中，我们使用预先训练过的 31 种经典卷积神经网络 (CNN)。在快速射电暴/非快速射电暴分类中，在第一个历元训练中达到 90-92% 的准确率，在真实数据测试中达到 99.8% 的最大准确率。

关键词　快速射电暴, 机器学习, 数据集

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

1　介绍

快速射电暴（FRB）是持续时间为毫秒或更短的明亮射电辐射脉冲[1,2]。自 2007 被发现以来，以 ASKAP、CHIME 和 FAST 为代表的国内外众多射电望远镜取得了系列观测进展和突破，推动了这一领域成为国际天文学新兴前沿热点。传统的快速射电暴搜寻使用消色散方法，首先从望远镜观测原始文件中读取"filterbank"[3] 或者 "Fits"[4] 文件，并去除射频干扰（RFI）[5,6]，然后需要搜索 100 到 2600 pc cm^{-3} 范围内的大量色散度量（DM）来寻找候选体[7,8]。

几乎所有的快速射电暴搜寻管线[9–12] 都采用传统的消色散算法进行盲搜索。尽管已经研发了很多优化算法[13–16]，然而此类算法仍然有一些缺点：大量的 DM 步骤会消耗大量的计算能力；太多的候选体需要人工确认；需要小心地去除射频干扰信号，否则会有太多假或者伪的结果。

我们生成了一套快速射电暴搜寻的机器学习数据集，用于在观测原始数据文件中检测快速射电暴。与在候选体中搜寻[17,18] 的方法不同，直接在观测的原始数据中搜索可以节省大量消色散的计算需求

和消除干扰信号的时间，以及检测弱快速射电暴信号 [19] 的可能性。此外，通过训练提高机器学习方法的准确度，也可以大幅减少最终候选体的数量。

为了开展机器学习搜寻快速射电暴，我们研发了 STEP 软件系统1)，机器学习的准确率与训练集有很大关系，而目前尚无快速射电暴的大型数据集，本文介绍了利用 ASKAP 开放数据创建数据集的方法。生成该数据集的主要目的是改进和优化在原始数据中搜寻快速射电暴的模型。数据集已有上万幅快速射电暴图像，由 STEP 在澳大利亚平方公里阵列探路者（ASKAP）[20] 公开的数据中检测到的 39 个 FRB 信号模拟产生。图1显示了 2 次 FRB 观测中检测到的 4 个 FRB 信号（在不同光束中）。该数据集将公开发布，供 FRB 科学界使用。

2 数据集构建

构建的快速射电暴数据集基于 ASKAP 开放的已知快速射电暴样本 [21]。构建步骤是首先使用传统的消色散管线检测原始数据中的所有已知快速射电暴。将这些消色散后的快速射电暴信号提取后，用以模拟快速射电暴信号集。最后通过随机选择原始背景数据、快速射电暴信号集和下面介绍的方法和参数选择，就可以构建面向原始数据的快速射电暴数据集（参见图2）。

2.1 快速射电暴观测和搜寻

ASKAP 的开放快速射电暴数据随附于论文 [21]。它发布了 19 次快速射电暴观测，每个都包含 36 波束的 “filterbank” 文件。数据为 8 比特，336 个 1 MHz 通道，采样时间为 1.26 毫秒，按下边带排序，最高频率为 1488 MHz2).

我们使用 STEP 来搜寻快速射电暴信号。这是我们团队自研的一个基于 GPU 的开源工具包，用于快速射电暴搜寻和分析。它在中国 SKA 区域中心原型机（CSKA-P）[22] 上进行了开发和测试，并从 ASKAP 公开的快速射电暴数据中搜寻出了所有已知快速射电暴信号，处理 36 波束、3295 秒数据的观测，使用单个 GPU 卡只需要 2743 秒。

2.2 模拟仿真快速射电暴信号

目前快速射电暴信号的特性还在不断发现和解析，所以模拟仿真快速射电暴信号最佳的方法是基于已有的真实快速射电暴信号。因此我们使用 STEP 搜寻并提取消色散后的快速射电暴信号，然后通过下述方法和参数仿真快速射电暴信号，最后将模拟仿真的快速射电暴信号注入真实的观测背景数据就可以生成数据集样本。

以下是影响数据集的几个因素及创建数据集的方法。

2.2.1 色散量

色散量是快速射电暴的主要特性，它决定了最高和最低频率之间的色散延迟时间。目前已知快速射电暴的色散量的范围是 100 到 2600pc cm^{-3} [7,8]，但是为了搜寻更远的奇异快速射电暴，搜寻的色散量范围越大越好，所以最大值可以超出 2600pc cm^{-3}；最小值 100pc cm^{-3} 一般认为是脉冲星或者射频干扰信号的色散范围所以维持不变。在我们的模拟仿真中，选择的快速射电暴色散量是从 100pc cm^{-3} 到由数据带宽、频率、采样时间和图像像素确定的最大值之间随机选择的，请参见下面的 4.1节。色散量对延迟时间是线性关系，所以对于超过图像最大值对应的色散量，可以通过对原始数据进行指定色散量的预处理来扩展对更大色散量的支持。比如图像最高支持色散量 400，超过 400 的色散量需要对数据进行 DM 为 300 的消色散的预处理，这里 100 色散量的差值是因为色散量小于 100 不在搜寻范围内。预处理后支持的最大色散量提高到 700。以此类推，对于更高色散量，只需对该数据继续做 DM 为 300 的消色散处理。基于这种方法，可以根据本地算力和具体需求自定义最小和最大色散量范围。应当注意的是，这种消色散预处理的方法只适

1) https://github.com/Xu-Zhijun/STEP

2) Shannon，Ryan；Bannister，Keith（2018）：Data from the ASKAP latitude 50 Fast Radio Burst (FRB) sample. v3. CSIRO. Data Collection. https://doi.org/10.25919/5b6ae6b515850

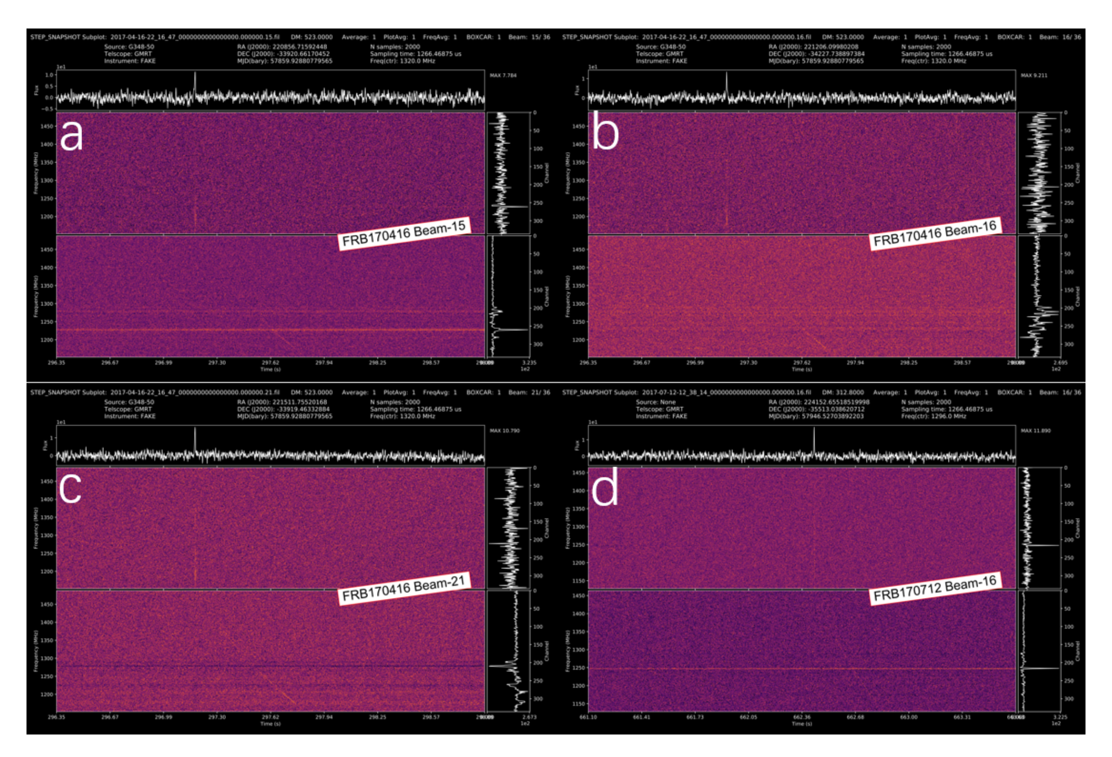


图 1　使用 STEP 搜寻到的快速射电暴的样本图像。图中（a）、（b）、（c）分别是 FRB170416 的第 15、16、21 波束。右下（d）为 FRB170721 的第 16 波束。在每个子图中，下方是原始数据图形，中上方是消色散后图像，顶部是去消色散数据按频率相加后的幅度分布。

Figure 1　Sample images of FRBs detected by STEP. Panel (a), (b), (c) are FRB170416 beam 15, 16, 21. Panel (d) is FRB170721 beam 16. In every panel, below is raw data, the middle is dedispersion data, and the top is the sum of the dedispersion data by frequency.

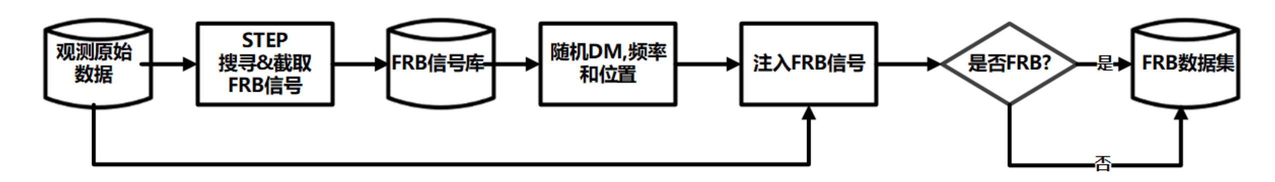


图 2　构建面向原始数据的快速射电暴机器学习数据集的流程图。首先从观测数据中检测 FRB，然后提取 FRB 信号来构建 FRB 数据集，最后采用随机 DM、流量和位置偏移的随机 FRB 信号注入原始背景数据中。

Figure 2　Flow chart explaining the processes to build MADFRB. We first detected the FRBs from observation data, and then extract the FRB signals to build the FRB dataset. The FRB signals then are randomly selected with random DM, fluence, and offset to inject to the raw data.

用于推理阶段，快速射电暴数据集和训练阶段是不需要的，这里涉及主要是解释如何在实际观测数据处理时如何扩展色散量范围。

2.2.2 射频干扰信号

仿真的快速射电暴信号最后被注入原始数据中，原始数据背景的射频干扰信号仍然会保留。ASKAP 的台址条件非常好，在大多数情况下，数据中射频干扰信号很少见，因此需要手动注入的方式模拟一些射频干扰信号。在我们的模拟仿真中，射频干扰信号被设置为色散量为零或者负数信号，这使得机器学习算法只能识别色散量为正且大于 100pc cm^{-3} 的信号，这个限制符合实际观测。

2.2.3 流量强度

观测发现快速射电暴的能量或流量变化很大，即使同一个快速射电暴比如重复的快速射电暴，不同脉冲的流量也不相同，甚至同一快速射电暴的同一脉冲，不同波束中的流量也不一样（见图1），这表明在适当范围内调整模拟仿真信号的流量强度不仅可以扩充样本数，还可以更接近真实情况。同时为了增强弱快速射电暴信号的搜寻能力，在我们的模拟仿真中，随机增强或者减弱信号流量，流量范围从接近检测上限到真实快速射电暴的信号流量。

2.3 数据集统计信息

ASKAP 的 FRB 观测数据有 36 个波束数据，有 FRB 信号的波束都是已知的，这样没有 FRB 信号的波束数据就可以作为噪声背景来生成无 FRB 的背景图像，同时通过注入 RFI 信号和 FRB 信号就可以生成 RFI 和 FRB 图像。目前数据集总共有 16040 个图像，包括 4010 个原始数据（无 FRB）、RFI、FRB 和弱 FRB 图像（参见图3）。我们用于模拟的真实 FRB 信号来自 FRB170906 的 4 个波束（见图 fig:FRB170906），我们使用的原始数据来自该观测 [21] 的其他没有快速射电暴信号的波束数据，这些数据也用于注入信号以模拟 FRB 或 RFI。

数据集的数据格式包括图像和 NumPy 文件。图像文件可以是 “ps”、“pdf”、“svg”，默认情况下也可以是卷积神经网络（CNN）算法支持的 “png” 格式。STEP 支持 Numpy “npy” 格式，或将其转换为其他 FRB 搜索管道支持的过滤器库文件。它还支持经典机器学习算法（如向量机、随机森林等）。因此，我们将有机会使用相同的数据集比较 CNN、经典 ML 算法和传统 FRB 管线的性能。

3 试验

试验在中国 SKA 区域中心原型机上进行 [23]。目前中国 SKA 区域中心原型机具备完善的软件平台 [24]，通过高速网络传输 SKA 先导设备数据 [25, 26]，并开展了射电天文管线的优化研究 [27, 28]。硬件设备拥有 4 个 GPU 节点，共包括 16 块英伟达 v100 显卡和 4 块 A40 显卡。本次试验通过 3 个 GPU 节点的 16 块英伟达 V100 显卡训练完成。

3.1 模型架构

我们使用 PyTorch3) [29] 及其自带的 torchvision 软件包一起开发用于测试快速射电暴数据集的模型。torchvision 软件包包括最新的可访问数据集、流行的模型架构和标准图像转换。在 PyTorch 1.6.0 版本中，模型软件包包含以下图像分类算法以及预训练模型：

- VGG (vgg11, vgg13, vgg16, vgg19, vgg11_bn, vgg13_bn, vgg16_bn, vgg19_bn) [30]
- DenseNet (densenet121, densenet169, densenet201, densenet161) [31]
- ResNet (resnet18, resnet34, resnet50, resnet101, resnet152) [32]
- ResNeXt (resnext50_32x4d, resnext101_32x8d) [33]
- Wide ResNet (wide_resnet50_2, wide_resnet101_2) [34]
- AlexNet [35]
- inception_v3 [36]
- GoogLeNet [37]
- mobilenet_v2 [38]

3) https://pytorch.org/

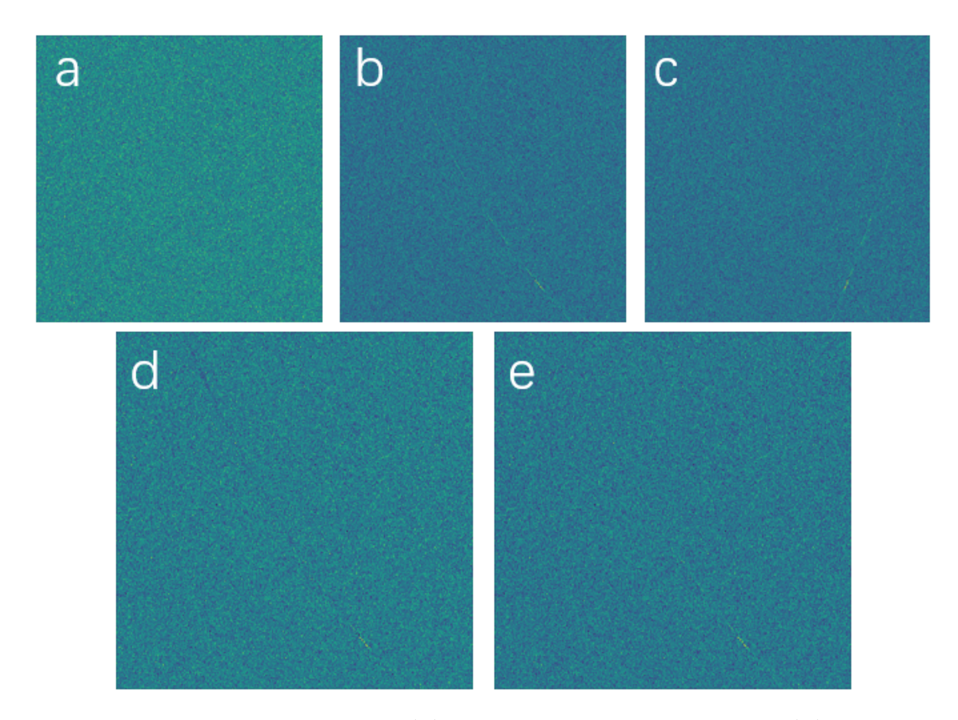


图 3　快速射电暴数据集中的图像。（a）是没有 FRB 信号的原始数据，（b）是真实搜寻到的 FRB 信号，（c）是射频干扰信号，（d）是具有错误弱信号的弱流量 FRB 信号，（e）是改正后正常的弱流量 FRB 信号

Figure 3　Images in the dataset. (a) the raw data without FRBs, (b) FRB with natural energy, (c) RFI, (d) weak FRB with wrongly weak signals, and (d) corrected weak FRB.

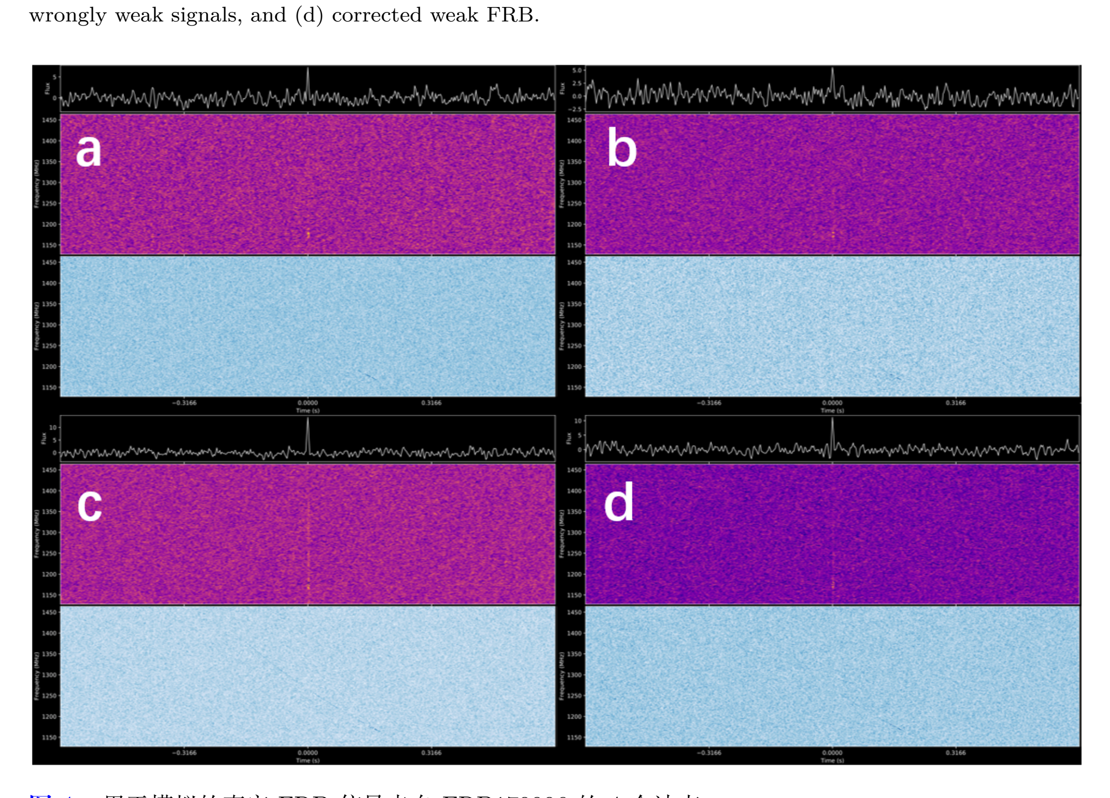


图 4　用于模拟的真实 FRB 信号来自 FRB170906 的 4 个波束

Figure 4　The real FRB signal used for simulation which come from four beams of FRB170906

- ShuffleNet v2 (shufflenet_v2_x0_5, shufflenet_v2_x1_0) [39]
- SqueezeNet (squeezenet1_0, squeezenet1_1) [40]
- MNASNet (mnasnet0_5, mnasnet1_0) [41]

3.2 预训练模型

常规的模型训练是用随机初始化来训练整个网络，这种方式耗时长，训练效果也很不稳定。本研究采用了使用预训练模型的方法。预训练模型是在一个巨大的数据集 [42] 上训练一个网络，然后将其用作初始化。由于我们的数据集可能与原始数据集非常不同，而且我们的数据集也可能很大，因此使用预训练模型后还需要通过训练对整个网络进行微调。

3.3 训练

我们以 8:2 的比例将数据集分成 12832 个图像训练集和 3208 个验证集。如上所述，本研究选取了含预训练模型的 31 种经典 CNN 算法对数据集进行试验。由于使用了预训练模型，在训练的第一个历元，大多数网络的准确率可以达到 90%，训练经过几次历元后，多数网络的准确率都超过了 99.7%。这说明经典 CNN 算法可以有效提取快速射电暴仿真数据集的图像特征，并通过多次迭代训练提高准确率。为了更好的研究不同 CNN 模型的分类效果，我们还计算了召回率 (Recall)、精确率 (Precision) 和 F1，其中召回率反应了模型识别正样本的能力，精确率反应了出现假阳性的概率，而 F1 反应了召回率和精确率在综合数值。从表2中可以看到有大多数模型的召回率、精确率和 F1 都在 99% 以上，表明我们的快速射电暴仿真数据集非常适合经典 CNN 类的算法。

3.4 测试数据集

为了验证基于快速射电暴仿真数据集训练后的 CNN 模型对于真实数据的分类效果，我们生成了 16544 张所有 ASKAP 公开的快速射电暴观测的真实 FRB 信号图像作为测试数据集，包括 35 张包含快速射电暴信号的图像，其他图像作为非快速射电暴图像。该数据截取了包含快速射电暴信号的约 10 秒原始数据，符合真实观测时的连续处理或者处理暂现源终端记录的候选体的缓存数据段。从表2中可以看到大多数网络的准确率可以达到 99%，而且层数更深的网络一般有更高的分类准确率。在实际应用中，为了尽可能减少错过正样本也就是快速射电暴信号，需要召回率尽量高。从表2中可以看到有 3 个模型的召回率都达到了 100%，表示所有的快速射电暴信号都找到了。召回率 100% 的原因是正样本的快速射电暴信号在真实数据中占比非常小，所以在正样本有限情况下召回率结果不一定精确，这也说明了真实场景下提高召回率的重要性。表2中，召回率 100% 情况下，精确率最高只有 99.321%，说明有一定几率出现假阳性候选体，需要一定的后续人工审查。

4 讨论

4.1 真实数据的图像大小和信号位置

快速射电暴数据集的首要限制就是图像的大小。卷积神经网络算法通常先将图像统一为固定大小，通常为等长的正方形图像。但在真实快速射电暴数据中，图像的高度由数据的通道数来定义，宽度由样本时间长度定义。观测中的通道数通常是固定的，因此对于真实数据我们需要选择与通道数等长的时间样本来获得等长图像。

目前数据集的生成算法中，我们默认快速射电暴的所有频段信号都在同一张图内。在处理真实快速射电暴数据时，为了保证这一点，需要对信号位置和色散量都有限制。对于未知的快速射电暴信号，为满足信号的位置限制，我们使用了设置重叠区域的方法。目前我们对 ASKAP 数据设置了 50% 重叠，比如单张图片包括时间长度为 1 的话，每个图片每经过 0.5 就自动生成一张图片。通过计算，这样可以保证在一定色散范围内所有频段信号都在单张图片内。

对于色散量限制，正如我们在 2.2.1节中提到的，色散量与延迟成正比，因此色散量的最大值也

表 1 31 种经典 CNN 模型的验证集结果

Table 1 31 classic CNN models with their respective validation result (Val Acc)

模型	验证集准确率 (%)	精确率 (%)	召回率 (%)	F1
resnet18	99.87	99.913	99.957	99.935
resnet34	99.731	99.896	99.835	99.865
resnet50	99.835	99.896	99.939	99.918
resnet101	98.986	99.904	99.079	99.49
resnet152	96.75	99.901	96.838	98.346
vgg11	99.818	99.913	99.904	99.909
vgg13	99.766	99.93	99.835	99.883
vgg16	99.783	99.904	99.878	99.891
vgg19	99.887	99.905	99.983	99.944
vgg11_bn	99.61	99.904	99.705	99.804
vgg13_bn	99.853	99.887	99.965	99.926
vgg16_bn	99.558	99.904	99.653	99.778
vgg19_bn	99.801	99.922	99.878	99.9
inception_v3	99.861	99.939	99.922	99.931
densenet121	99.827	99.948	99.878	99.913
densenet161	99.783	99.904	99.878	99.891
densenet169	99.879	99.939	99.939	99.939
densenet201	99.775	99.957	99.818	99.887
resnext50_32x4d	99.792	99.93	99.861	99.896
resnext101_32x8d	98.726	99.93	98.793	99.358
wide_resnet50_2	98.891	99.895	98.992	99.441
wide_resnet101_2	99.827	99.939	99.887	99.913
googlenet	99.818	99.93	99.887	99.909
mobilenet_v2	99.792	99.913	99.878	99.896
alexnet	99.827	99.922	99.904	99.913
squeezenet1_0	99.567	99.913	99.653	99.783
squeezenet1_1	98.977	99.93	99.045	99.485
mnasnet1_0	99.879	99.887	99.991	99.939
mnasnet0_5	99.307	99.774	99.531	99.652
shufflenet_v2_x0_5	99.766	99.904	99.861	99.883
shufflenet_v2_x1_0	99.567	99.913	99.653	99.783

表 2　31 种经典 CNN 模型的测试集结果
Table 2　31 classic CNN models with their respective test set result

模型	测试集准确率 (%)	精确率 (%)	召回率 (%)	F1
resnet18	99.127	98.344	99.938	99.134
resnet34	99.608	99.226	99.991	99.607
resnet50	98.896	97.862	99.982	98.911
resnet101	99.791	99.618	99.964	99.791
resnet152	99.194	98.901	99.492	99.196
vgg11	99.506	99.123	99.893	99.506
vgg13	99.581	99.247	99.92	99.582
vgg16	99.43	98.88	99.991	99.432
vgg19	97.885	95.952	100	97.934
vgg11_bn	99.657	99.321	100	99.659
vgg13_bn	98.642	97.398	99.946	98.656
vgg16_bn	99.657	99.317	100	99.657
vgg19_bn	99.069	98.206	99.982	99.086
inception_v3	97.217	94.951	99.761	97.297
densenet121	99.884	99.857	99.911	99.884
densenet161	99.813	99.665	99.965	99.814
densenet169	99.884	99.839	99.929	99.884
densenet201	99.817	99.731	99.901	99.816
resnext50_32x4d	99.532	99.133	99.938	99.534
resnext101_32x8d	98.753	99.403	98.091	98.743
wide_resnet50_2	99.1	98.3	99.919	99.103
wide_resnet101_2	99.706	99.45	99.964	99.707
googlenet	99.181	98.476	99.902	99.184
mobilenet_v2	99.693	99.441	99.946	99.693
alexnet	93.886	89.437	99.519	94.209
squeezenet1_0	99.212	98.814	99.616	99.213
squeezenet1_1	99.221	98.769	99.689	99.227
mnasnet1_0	96.277	93.201	99.876	96.423
mnasnet0_5	50.109	50.076	99.822	66.694
shufflenet_v2_x0_5	92.95	88.43	98.891	93.369
shufflenet_v2_x1_0	95.827	93.56	98.39	95.914

限制在给定时间长度的样本中。这个问题可以通过对数据在时域进行下采样来解决，但会带来灵敏度降低的风险。为了解决这个问题，我们采用了对高色散量信号进行预先消色散的解决方法。

目前 ASKAP 数据中我们设置了 400pc cm^{-3} 的预消色散步骤，比如对于搜索范围 2000pc cm^{-3}，需要设置 5 个并行搜寻管线，分别对应不进行预消色散和预消色散 400、800、1200、1600pc cm^{-3}。这样对于每个搜寻管线，只需要搜寻 400pc cm^{-3} 以内，所有信号都在单张图片的情况，有效提高搜索准确率。当然为了彻底解决限制问题，有必要解除快速射电暴所有信号都在单张图像中的限制，但这将带来其他问题，需要在后续研究中解决。

4.2 弱快速射电暴信号

该问题来源于本研究初期在进行随机流量强度变换时的错误。为了注入随机流量强度的快速射电暴信号，我们初期通过随机选择的流量强度因子对信号的所有频率流量同时进行增强或削弱。。检查生成的图像后，我们发现一些在真实数据中不会出现的 "奇异的弱线" (参见图3)。这个问题的原因是快速射电暴信号在不同频率上流量强度表现不同，对信号进行随机流量强度变换时，原先流量强度较低的频率就可能发生信号比背景的流量强度还弱的现象，造成了 "奇异的弱线"。为了解决这个问题，我们在仿真中对注入的快速射电暴信号进行检查，并将信号中的低于同频率背景强度的部分替换为背景噪声，这样仿真图像就更符合真实数据的情况。

4.3 快速射电暴宽度

快速射电暴信号的宽度决定了需要从观测数据中提取多少真实 FRB 信号样本。FRB 的最大和最小宽度仍不确定，因此必须手动提取真实的 FRB 信号。类似地，尽管 FRB 的宽度可以设置为随机选择，但在我们的模拟中，它默认设置为不变。

5 未来工作

数据集只是快速射电暴搜寻的第一步，未来还有很多工作需要继续。比如引入不同望远镜的观测和射频干扰数据以提高适应性。针对平方公里阵列第一期的快速射电暴研究，可以利用在 SKA 低频站址的 SKA 先导设备 MWA 数据，进行 SKA 低频快速射电暴仿真数据集研究。同样可以利用 MeerKAT 数据，研究 SKA 中频快速射电暴仿真数据集。另外快速射电暴的能量强度在不同频率会有一定随机性，不同数据格式的量化效果等还需要根据不同观测数据进行具体研究。目前快速射电暴搜寻管线研发的主要困难之一就是缺乏统一的比较手段。基于已有快速射电暴仿真数据集，进一步可以开展快速射电暴搜寻管线的比较和优化。可以有效量化管线性能，为快速射电暴搜寻管线研发、测试和优化提供统一标准。

6 结论

为了加速快速射电暴搜寻管线研究，我们开发了一套基于机器学习的快速射电暴数据集，它可以训练机器学习算法以搜寻原始数据中的快速射电暴。同时数据集也可以作为传统快速射电暴管线的性能量化标准。数据集目前已有 8020 个快速射电暴、4010 个非快速射电暴和 4010 个射频干扰图像，这些图像是根据公开的快速射电暴观测结果构建的。我们为最先进的人工智能算法提供开源数据集，以比较快速射电暴识别算法。该数据集为卷积神经网络和经典机器学习算法提供图像和 numpy 格式的文件。数据集可以实现快速射电暴/非快速射电暴分类，或快速射电暴/射频干扰/背景噪声分类。目前图像结果已经开源，下一步测试完成后，仿真的工具包也会开源，满足特定望远镜观测数据搜寻的需求。

致谢

本研究使用了中国 SKA 区域中心原型机的资源。

参考文献

1 Lorimer D R, Bailes M, McLaughlin, M A, et al. A Bright Millisecond Radio Burst of Extragalactic Origin. Science, 2007, 318(5851): 777–780

2 Petroff E, Hessels J W T, Lorimer D R. Fast radio bursts, Astron Astrophys Rev, 27(1): 1–75

3 Lorimer D R. SIGPROC: pulsar signal processing programs. Astrophysics Source Code Library, 2011, ascl: 1107.016

4 Hotan A W, Van Straten W, Manchester R N. PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis. Publ Astron Soc Aust, 2004, 21(3): 302–309

5 Petroff E, Keane E F, Barr E D, et al. Identifying the source of Perytons at the parkes radio telescope. Mon Notices Royal Astron Soc, 2015, 451(4): 3933–3940

6 Nita G M, Gary D E. The generalized spectral kurtosis estimator. Mon Notices Royal Astron Soc, 2010, 406(1): L60–L64

7 Bhandari S, Keane E F, Barr E D, et al. The SUrvey for Pulsars and Extragalactic Radio Bursts–II. New FRB discoveries and their follow-up. Mon Notices Royal Astron Soc, 2018, 475(2): 1427–1446

8 Caleb M, Keane E F, Van Straten W, et al. The SUrvey for Pulsars and Extragalactic Radio Bursts–III. Polarization properties of FRBs 160102 and 151230. Mon Notices Royal Astron Soc, 2018, 478(2): 2046–2055

9 Ransom S. PRESTO: PulsaR Exploration and Search TOolkit. Astrophysics source code library, 2011, ascl: 1107.017

10 Champion D J, Petroff E, Kramer M, et al. Five new fast radio bursts from the HTRU high-latitude survey at Parkes: first evidence for two-component bursts. Mon Notices Royal Astron Soc, 2016, 460(1): L30–L34

11 Bannister K W, Shannon R M, Macquart J P, et al. The detection of an extremely bright fast radio burst in a phased array feed survey. Astrophys J Lett, 2017, 841(1): L12

12 Mikhailov K, Sclocco A. The Apertif Monitor for Bursts Encountered in Real-time (AMBER) auto-tuning optimization with genetic algorithms. Astronomy and computing, 2018, 25: 139–148.

13 Barsdell B R, Bailes M, Barnes D G, et al. Accelerating incoherent dedispersion. Mon Notices Royal Astron Soc, 2012, 422(1): 379–392

14 Sclocco A, van Leeuwen J, Bal H E, et al. Real-time dedispersion for fast radio transient surveys, using auto tuning on many-core accelerators. Astronomy and Computing, 2016, 14: 1–7

15 Zackay B, Ofek E O. An accurate and efficient algorithm for detection of radio bursts with an unknown dispersion measure, for single-dish telescopes and interferometers. Astrophys J, 2017, 835(1): 11

16 Amiri M, Bandura K, Berger P, et al. The CHIME fast radio burst project: system overview. Astrophys J, 2018, 863(1): 48

17 Connor L, van Leeuwen J. Applying deep learning to fast radio burst classification. Astron J, 2018, 156(6): 256

18 Agarwal D, Aggarwal K, Burke-Spolaor S, et al. FETCH: A deep-learning based classifier for fast transient classification. Mon Notices Royal Astron Soc, 2020, 497(2): 1661–1674

19 Zhang Y G, Gajjar V, Foster G, et al. Fast radio burst 121102 pulse detection and periodicity: a machine learning approach. Astrophys J, 2018, 866(2): 149

20 McConnell D, Allison J R, Bannister K, et al. The Australian Square Kilometre Array Pathfinder: Performance of the Boolardy Engineering Test Array. Publ Astron Soc Aust, 2016, 33

21 Shannon R M, Macquart J P, Bannister K W, et al. The dispersion–brightness relation for fast radio bursts from a wide-field survey. Nature, 2018, 562(7727): 386–390.

22 An T, Wu X P, Hong X. SKA data take centre stage in China. Nat Astron, 2019, 3(11): 1030–1030

23 An T, Wu X C, Lao B Q, et al. Status and progress of China SKA Regional Centre prototype. arxiv:2206.13022

24 Lao B Q, Zhang Y K, An T, et al. Software Platform on China SKA Regional Center Prototype System(in Chinese).ChinaXiv:202206.00173. [劳保强, 张迎康, 安涛, 等.(2022). 中国 SKA 区域中心原型系统 – 软件平台.ChinaXiv:202206.00173]

25 Guo S G, An T, Xu Z J, et al.Progress and Prospect of transcontinental high-speed data transmission at SKA Regional Center in China(in Chinese). ChinaXiv: xxxx [郭绍光, 安涛, 徐志骏, 等.(2022). 中国 SKA 区域中心跨洲际高速数据传输进展及展望.ChinaXiv:xxxx]

26 Guo S G, Lu Y, An T, et al. Scientific data flow and array simulation analysis for the SKA-1 era(in Chinese). ChinaXiv:xxxx. [郭绍光, 安涛, 徐志骏, 等.(2022). 面向 SKA-1 时代的科学数据流及阵列模拟分析.ChinaXiv:xxxx]

27 Wei J W, Zhang C F, Zhang Z L, et al. Parallel optimization of the pulsar search pipeline on China SKA Regional Centre

Prototype (in Chinese). ChinaXiv:T202206.00297 [韦建文, 张晨飞, 张仲莉, 等.(2022). 低频射电脉冲星搜索的性能优化方法. ChinaXiv:T202206.00297]

28 Wei J W, Zhang C F, Lao B Q, et al. Optimization of parallel processing of Square Kilometre Array low frequency imaging pipeline (in Chinese). ChinaXiv:T202206.00292.[韦建文, 张晨飞, 劳保强, 等.(2022).SKA 低频成像管线并行优化化.ChinaXiv:T202206.00292]

29 Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019, 32: 8026-8037

30 Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis, 2015, 115(3): 211–252.

31 Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708.

32 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778

33 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492–1500

34 Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint, 2016, arXiv:1605.07146

35 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25: 1097–1105

36 Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818–2826

37 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1–9.

38 Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, 2017, arXiv:1704.04861

39 Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848–6856

40 Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint, 2016, arXiv:1602.07360

41 Tan M, Chen B, Pang R, et al. Mnasnet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2820–2828

42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014, arXiv:1409.1556

A machine learning dataset for FRB detection in raw data

XU ZhiJun[1,2*], AN Tao[1,2], GUO ShaoGuang[1,2], LAO BaoQiang[1,2], LU WeiJia[1,2] & WU Xiaocong[1,2]

1. SKA Regional Centre Joint Lab, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China;
2. SKA Regional Centre Joint Lab, Peng Cheng Lab, Shenzhen, 518066, China

We introduce a machine learning FRB dataset that can train the ML algorithms to reach the FRBs in raw data. It has 8020 FRB simulation images, 4010 non-FRB and 4010 RFI simulation images built from the public FRB observations, and can be expanded in any number as needed. This work provides an open-source dataset for state of art AI to the comparison of FRB event recognition algorithms. The dataset provides image and NumPy format files for both convolutional neural networks and classic machine learning algorithms. The dataset can implement FRB/non-FRB classification, or FRB/RFI/Blank classification. In the example, we used 31 pre-trained classic CNNs. In FRB/non-FRB classification, it achieves the accuracy of 90-92% in the first training epoch and max accuracy of 99.8% in real FRB dataset testing.

FRB, Machine Learning, Dataset

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

doi: